Text-Emoji Correspondenece

AI Final Project Team12

111550024 蔡芳慈、111550168 陳奕、111550113 謝詠晴、111550155 郭芷杼

1. Introduction

- Emojis are essential for conveying emotions and context in digital communication
- Chinese text-to-emoji prediction system
- We will analyze the performance of the system using different numbers of classes and models

今天忙了一整天

終於把期末專題完成

我真的很棒

今天忙了一整天

終於把期末專題完成

我真的很棒

2. Related work

- Other Work
  - Used the distillbert-based-uncased model to predict
  - No other data processing except tokenization
  - Have a similar input format which are in English
  - Some of them have an output format containing fixed multiple emojis
    - lyrics to emoji : https://www.kaggle.com/code/aguschin/lyrics-to-emoji

2. Related work

- Other Work
  - Some of them allow the user to specify the number of output emojis
    - Text to emoji : https://github.com/andylolu2/Text2Emoji
- Our Work
  - Input is in Chinese
  - Output only one emoji for a single sentence.

3. Dataset

- Web Crawler
  - Instagram (instaloader)
- Data Collection
  - 22 accounts
  - 2000 records per account
    - article content
    - 20 comments per account

3. Dataset

- Data Preprocessing

  input : @jaychou蘋果肉桂waffle一份竟然$204！

  preprocess : ~~@jaychou~~ 蘋果 肉桂 ~~waffle~~ 一份 ~~竟然~~ ~~$~~ ~~204~~ ~~!~~

  output : “蘋果”,”肉桂“,”一份“

- Emoji Analysis
  - The top 20, 30, and 40 most commonly used emojis

3. Dataset

- Statistics
  - Total collected data : 41273
  - Preprocessed data : 35507
  - Final data set (most commonly used emojis) :
    - Top 20 : 20078
    - Top 30 : 22783
    - Top 40 : 24596

4. Baseline

- Dataset
  - Only the top 20 most commonly used emojis
  - Contains 50,000 English tweets from X (formerly known as Twitter)
  - Each line contains one or more emojis
  - Translate the data into Chinese as our baseline dataset

4. Baseline

- Model
  - Bidirectional Long Short-Term Memory (BLSTM)
- Trained
  - take Chinese text as input
  - accurately predict the most appropriate emoji from a predefined set of 20 commonly used emojis.

4. Baseline

翻譯 baseline dataset 結果

```
df.head(10)
```

	Tweet	Label
0	小復古最愛的人水牆	0
1	華麗昨天 kcon 化妝 羽毛	7
2	民主廣場喚醒令人震驚的結果決定 NBC 新聞	11
3	amp vilo 華特迪士尼魔法王國	0
4	銀河系很遠很遠	2
5	今晚佛羅裡達晚餐 煎鮭魚 蒸粗麥粉 蔬菜沙拉 美味的晚餐 佛羅裡達鮭魚	1
6	最喜歡的高級比賽恭喜擊敗西西塞勒姆	8
7	得到了正式的最好的朋友 phi mu jsu	0
8	原因想念小兄弟復古表弟愛印第安納大學	13
9	生日吻麥迪遜威斯康辛州	9

accuracy: 0.9357

0	❤️
1	😍
2	😂
3	💕
4	🔥
5	😊
6	😎
7	✨
8	💙
9	😘
10	🇺🇸
11	us
12	🌞
13	💜
14	😉
15	💯
16	😁
17	🎄
18	🇺🇸
19	😜

```
Enter tweet
今天太陽真大好熱

Emojified Tweet

1/1 [==============================] - 0s 20ms/step
今天太陽真大好熱 🌞
```

```
Enter tweet
開始聖誕假期

Emojified Tweet

1/1 [==============================] - 0s 44ms/step
開始聖誕假期 🎄
```

5. Approach

- Data Preprocessing（Example）
  - Original : 白桃烏龍茶雪華夫餅 整份$204、半份$116
  - Preprocessing : ['白桃', '烏龍茶', '雪華夫餅', '整份', '半份']
  - Tokenize : [3536, 3980, 3981, 1854, 1855]
  - Padding :

```
[  0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0
       ......
   0   0   0   0   0   0   0  3536 3980 3981 1854 1855 ]
```

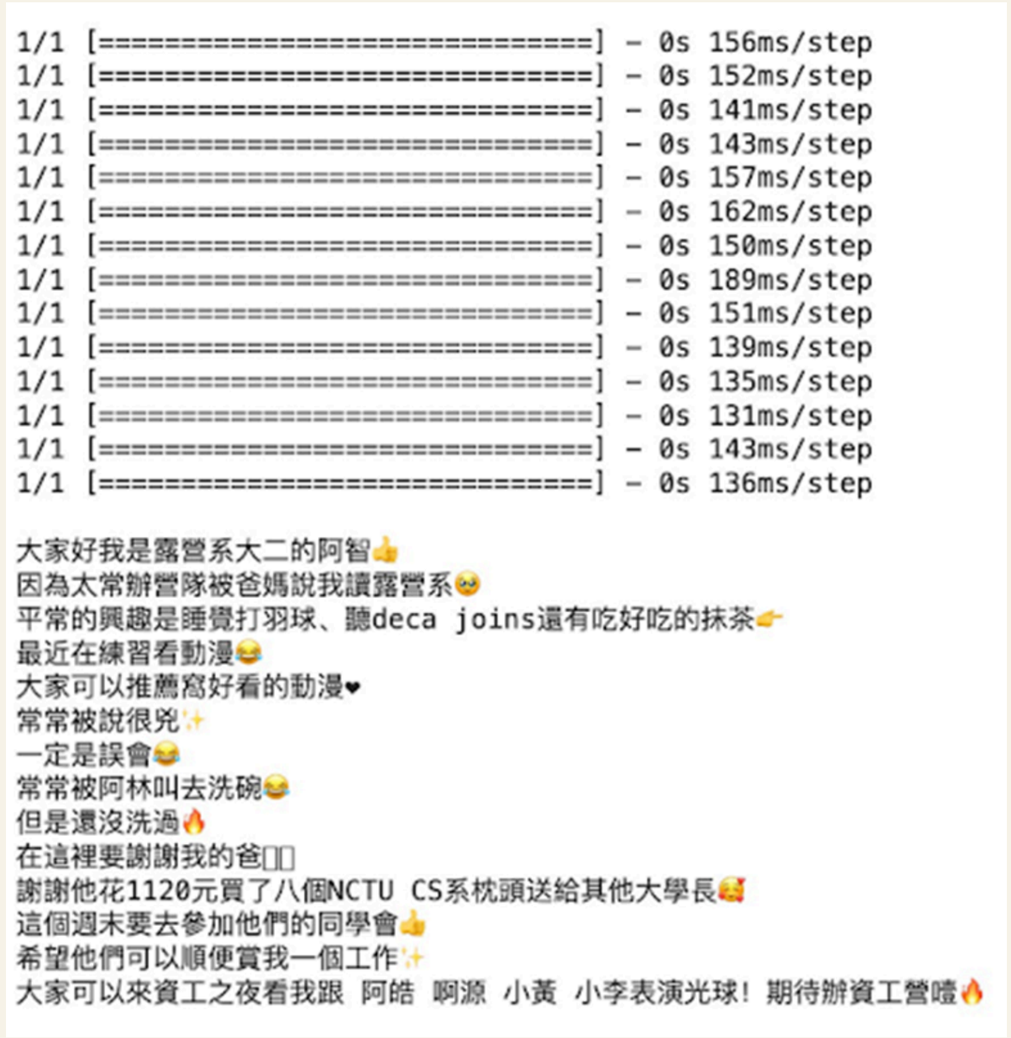
5. Approach

- Model Architecture
  - Embedding layer
  - BLSTM layer
  - Fully connected layer

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 140, 128)          2317056

 bidirectional (Bidirection  (None, 140, 160)          133760
 al)

 bidirectional_1 (Bidirecti  (None, 140, 160)          154240
 onal)

 global_max_pooling1d (Glob  (None, 160)               0
 alMaxPooling1D)

 dropout (Dropout)           (None, 160)               0

 dense (Dense)               (None, 64)                10304

 dropout_1 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 20)                1300

=================================================================
Total params: 2616660 (9.98 MB)
Trainable params: 2616660 (9.98 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

5. Approach

- Training Process
  - Loss function
  - Optimizer
  - Batch Training
- Prediction
  - Preprocessing
  - Model Inference



5. Approach


Emoji Transformer
歡迎使用 Emoji 轉換器。請輸入文本讓我們幫您加上表情符號吧～
輸入文本...
轉換
結果: 我是真的不能控制我自己 💪 每天想你 甚至會出現在夢裡 👉 把對你的情緒全部哼成旋律 😻 還是期待幻想的畫面會不會發生😻

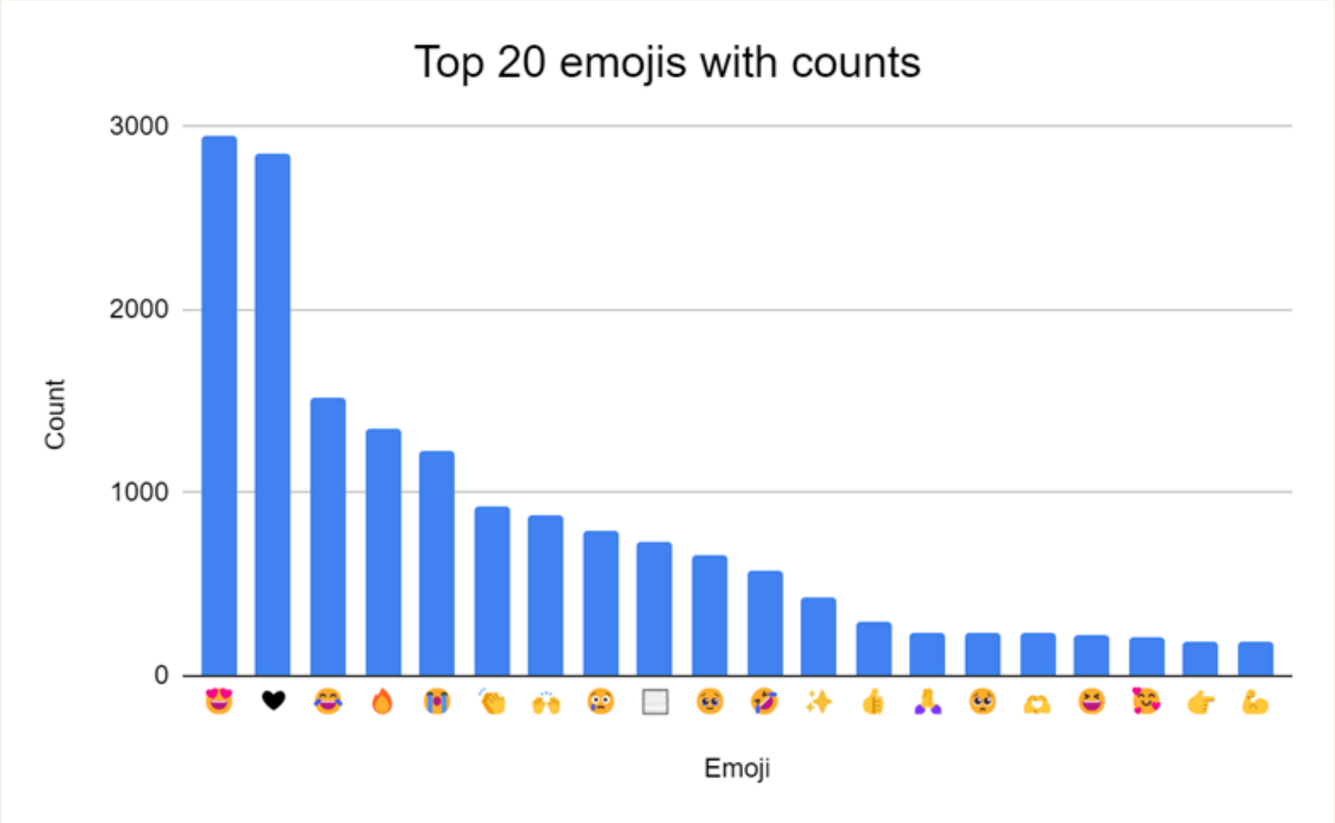
6. Evaluation

- Quantitative
  - f1-score of each class(emojis)
  - accuracy of total prediction
  - confusion matrix
- Qualitative
  - 易用性
  - 可解釋性
  - 用戶體驗

7. Results

（第一次採用的 dataset）

- Top 20 emojis



7. Results

(第一次採用的 dataset)

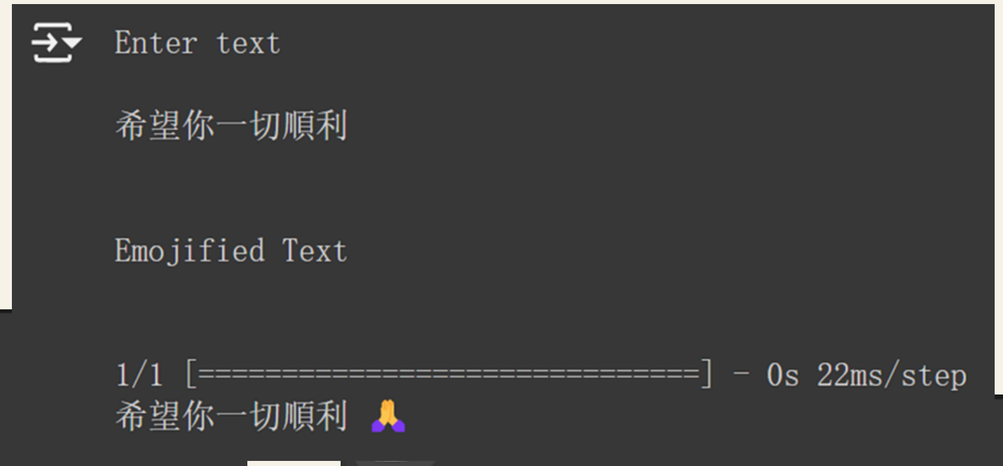
- BLSTM
- accuracy: 0.7263
- problem:
  - 訓練資料來源不平均

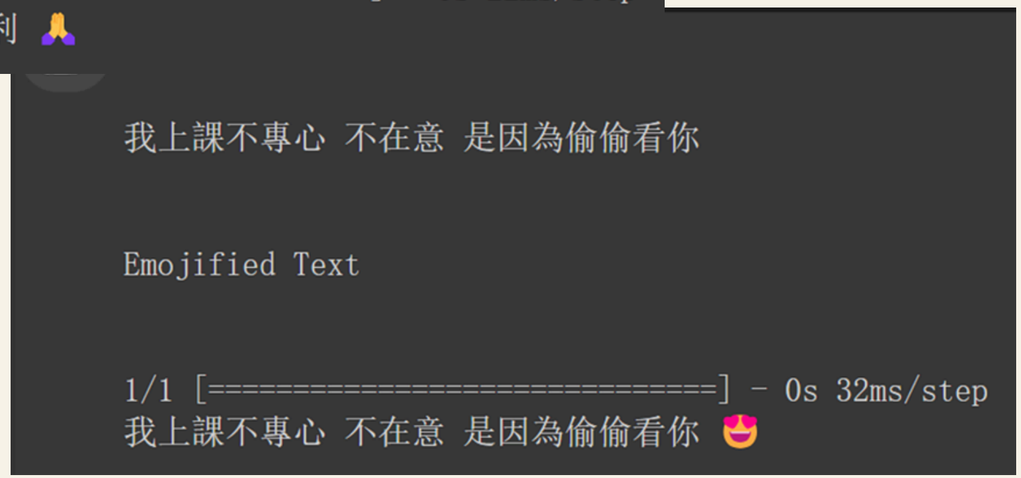
```
369/369 [==============================] - 23s 61ms/step
              precision    recall  f1-score   support

           0       0.44      0.17      0.25       605
           1       0.36      0.18      0.24       607
           2       0.80      0.66      0.73       586
           3       0.53      0.52      0.52       576
           4       0.66      0.67      0.67       610
           5       0.59      0.64      0.61       581
           6       0.55      0.60      0.57       569
           7       0.72      0.80      0.76       620
           8       0.71      0.29      0.41       581
           9       0.80      0.72      0.76       586
          10       0.81      0.92      0.86       545
          11       0.94      0.91      0.92       571
          12       0.72      0.78      0.75       606
          13       0.72      0.95      0.82       561
          14       0.82      0.91      0.86       578
          15       0.68      0.91      0.78       613
          16       0.85      0.96      0.90       614
          17       0.84      0.96      0.90       624
          18       0.89      1.00      0.94       561
          19       0.74      0.98      0.84       602

    accuracy                           0.73     11796
   macro avg       0.71      0.73      0.70     11796
weighted avg       0.71      0.73      0.70     11796
```

7. Results

(第一次採用的 dataset)

- Some example





7. Results

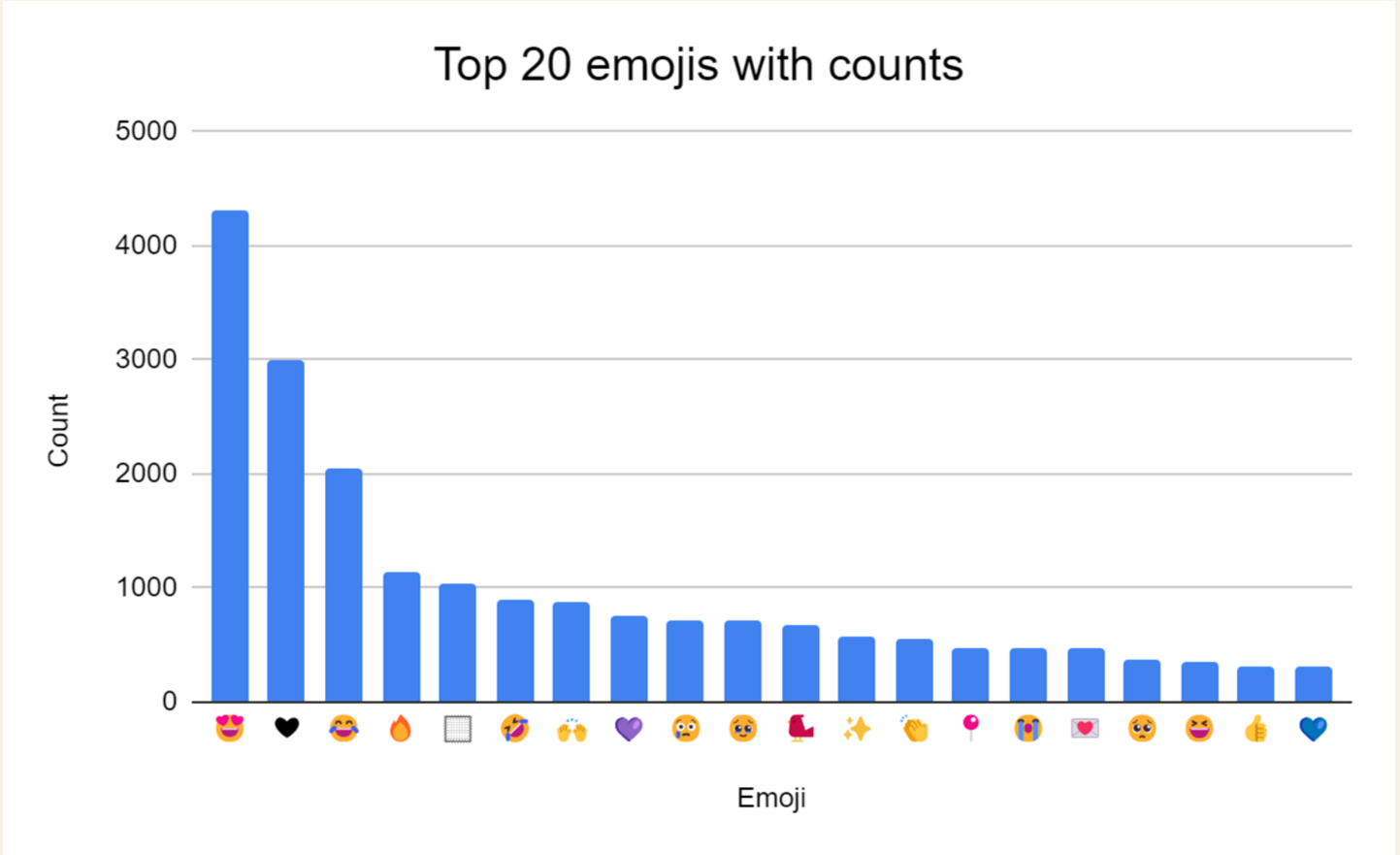
（第一次採用的 dataset）

- 在爬蟲時就只下載包含emoji的資料
  - 確保資料來源平均
- BLSTM較LSTM表現較好，因此採取BLSTM方法
- 分別用前20,30和40個常用emoji的資料去訓練model

7. Results

(最終採用的 dataset)

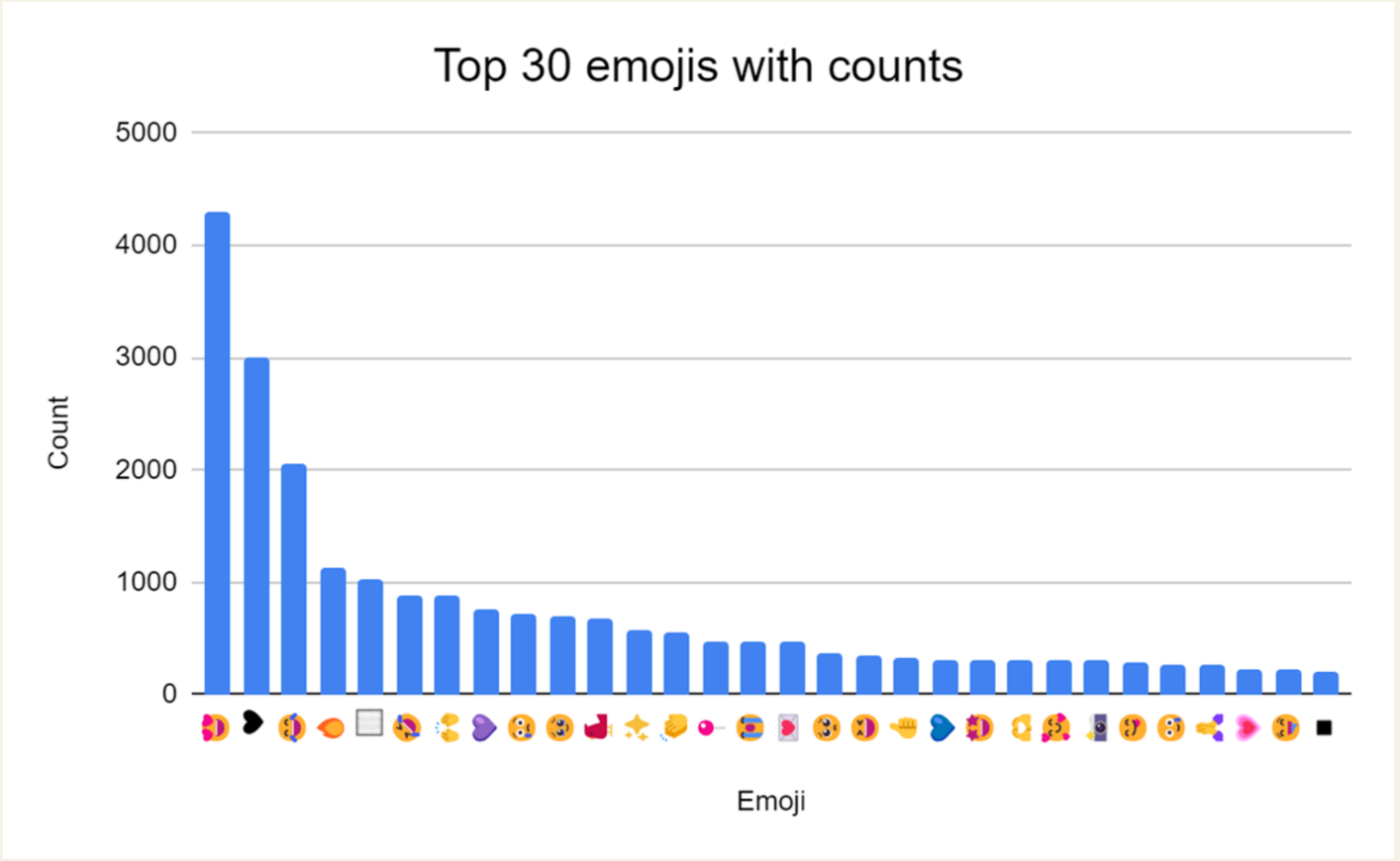
- Top 20 emojis



7. Results

(最終採用的 dataset)

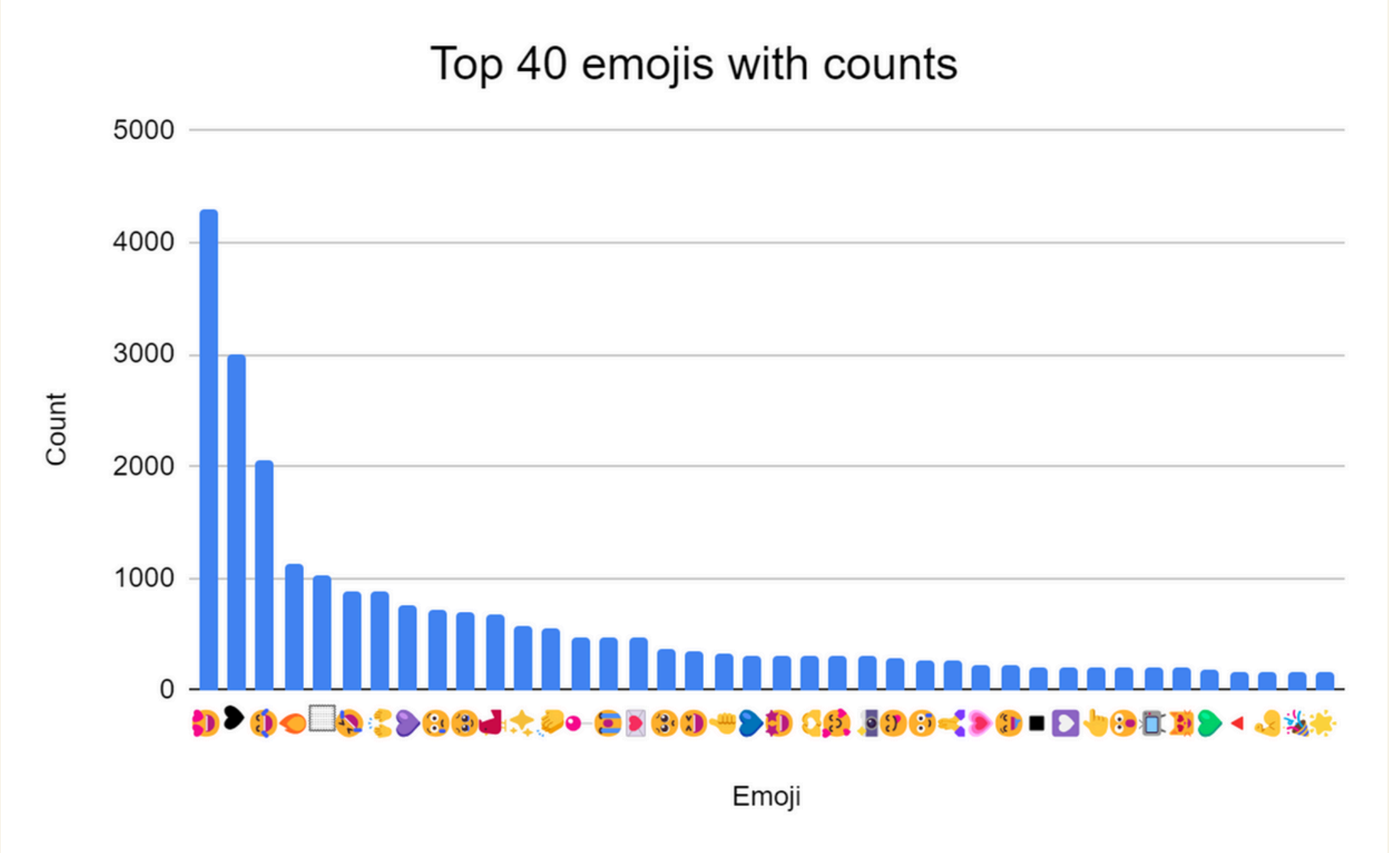
- Top 30 emojis



7. Results

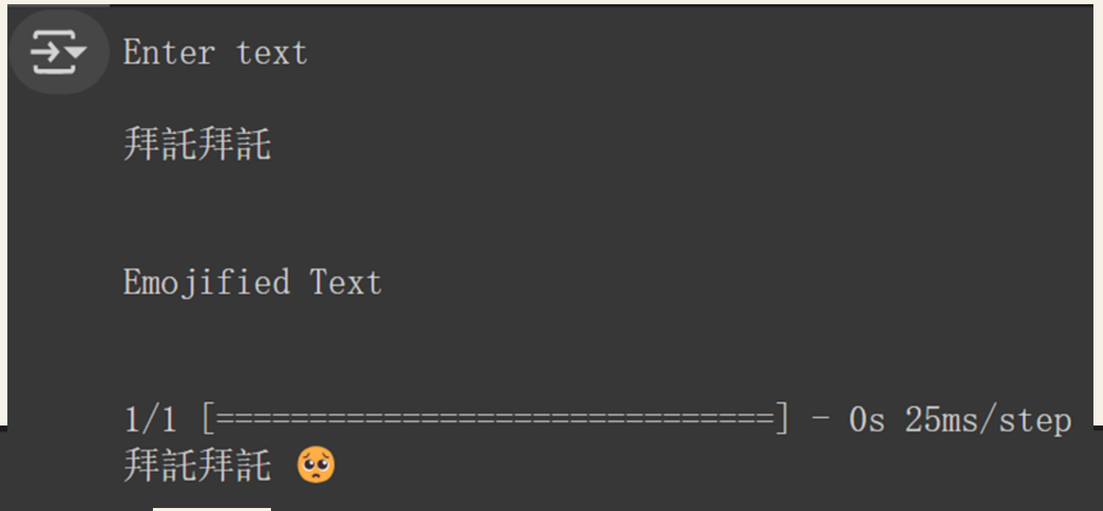
(最終採用的 dataset)

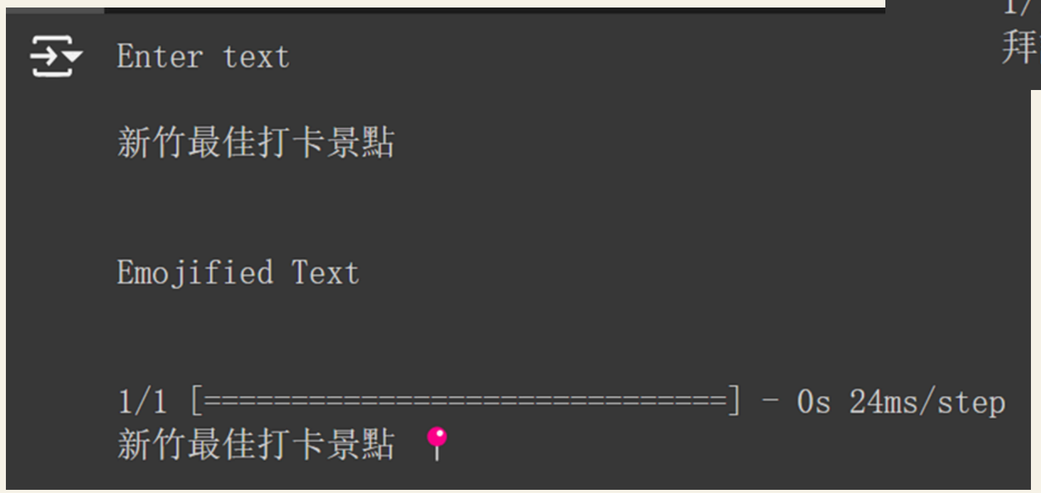
- Top 40 emojis

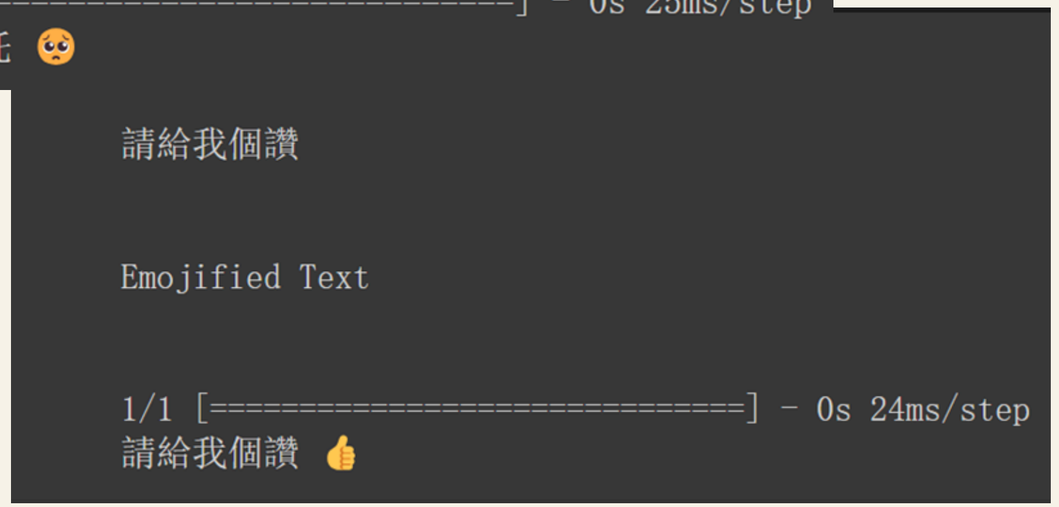


7. Results (最終採用的 dataset)

- Some example
  (30 emoji model)



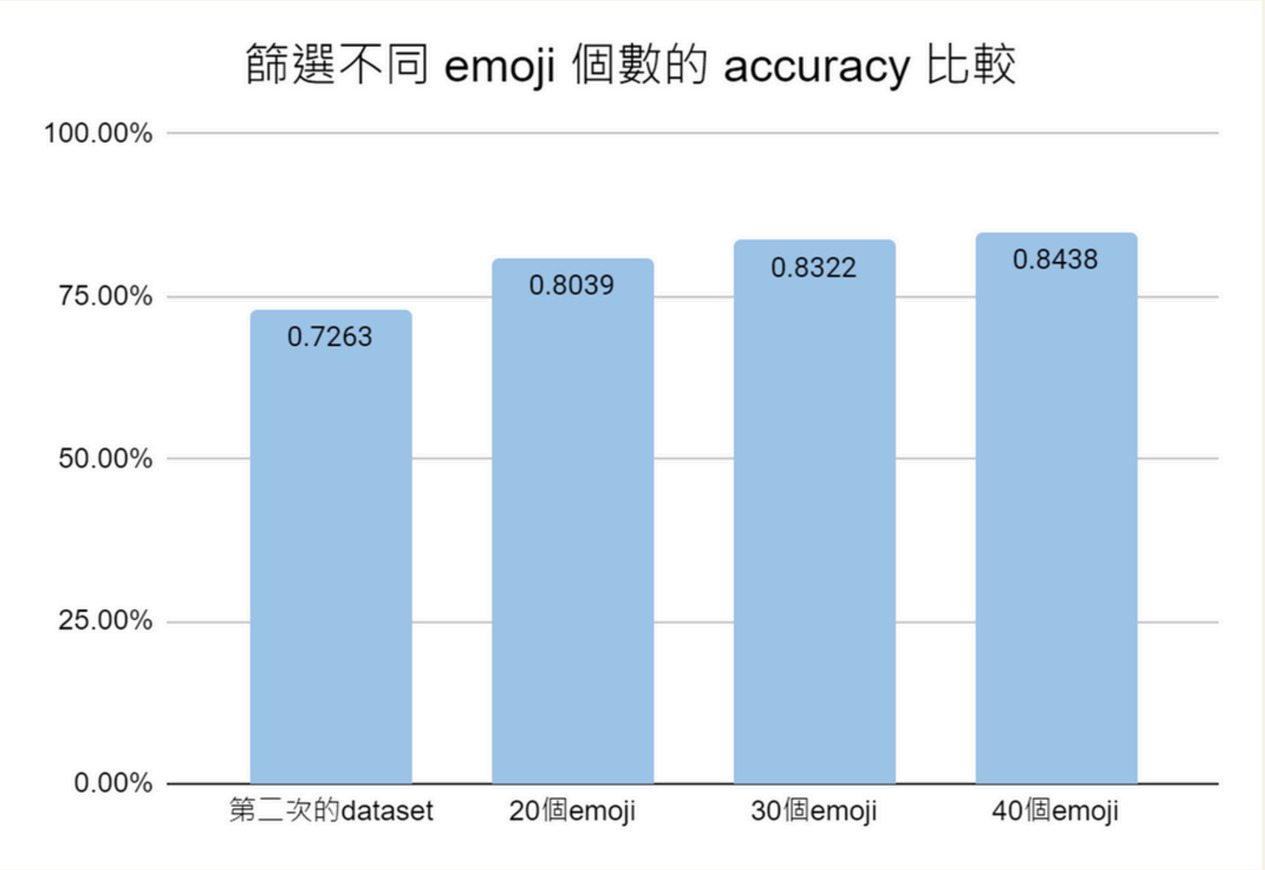




DEMO

8. Analysis

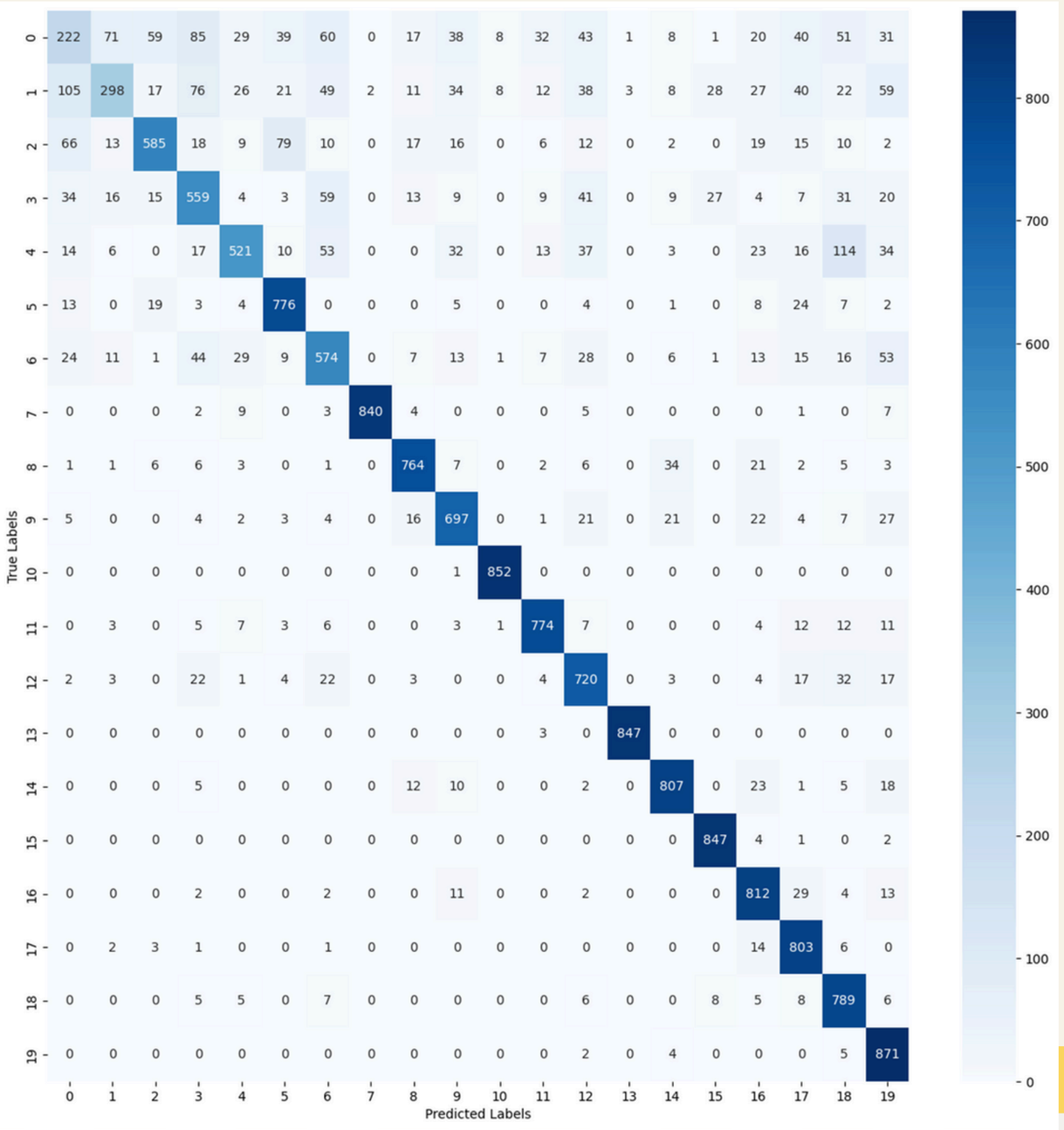
- accuracy comparison
- 有達到八成



8. Analysis

- confusion matrix

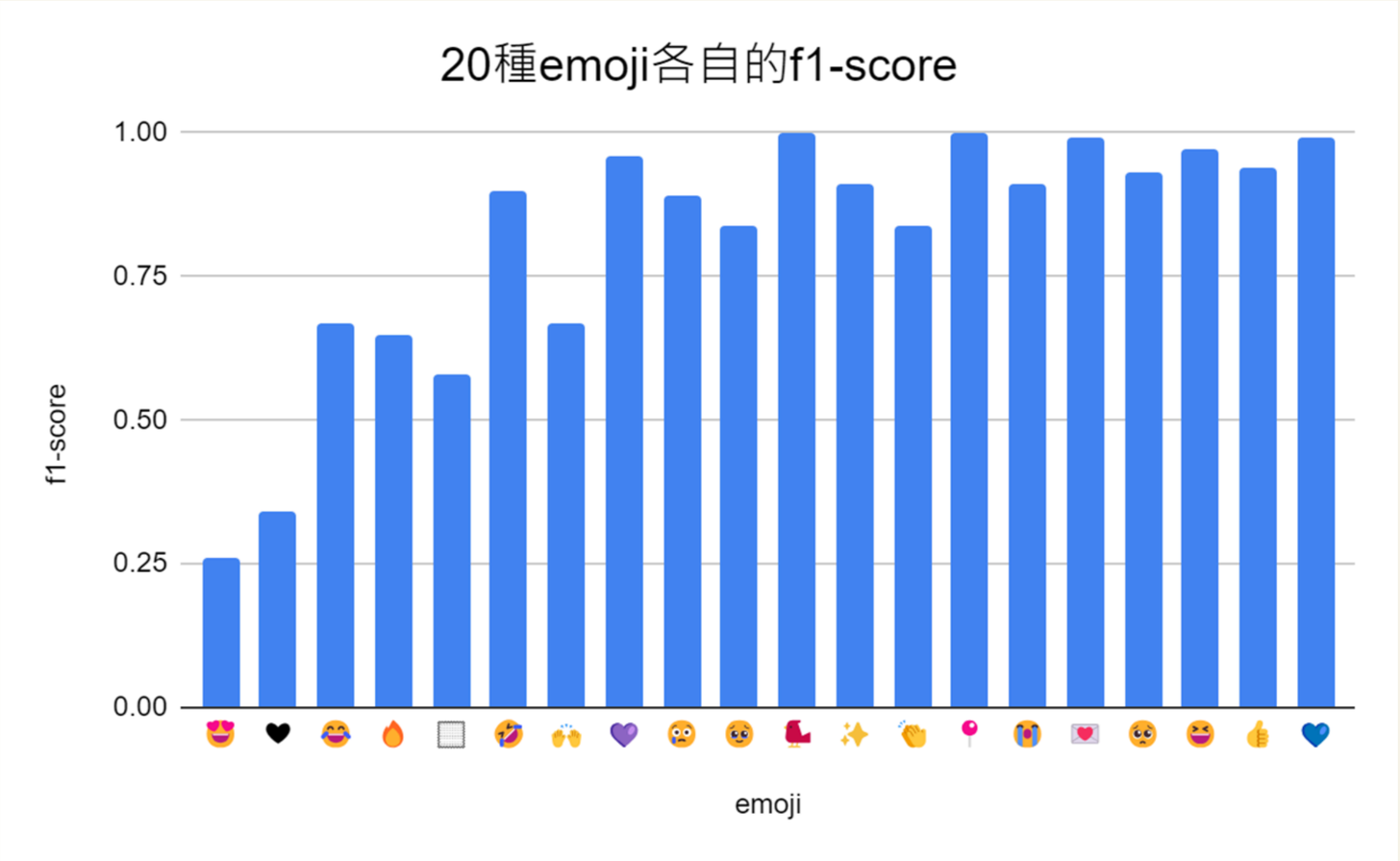
[最終採用的 dataset]



8. Analysis

[最終採用的 dataset]

- f1 score of each emoji



8. Analysis

- Error analysis

1.輸入只含有stopwords的句子

2.只能接受中文

3.仍有不合理的輸出

```
雖然

1/1 [==============================] - 0s 24ms/step
雖然😭
```

```
It's time to go to bed

1/1 [==============================] - 0s 25ms/step
It's time to go to bed😭
```

```
今天心情真美麗

1/1 [==============================] - 0s 24ms/step
今天心情真美麗😢
```

9. Future Work

- How to improve
  - Larger and more diverse data sets
  - Data preprocessing
  - Add emoji categories
  - Multiple emojis in one sentence
  - Instant feedback

9. Future Work

- Challenges and Solutions

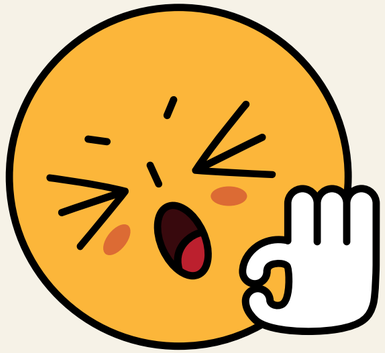
1. 不同使用者對文字解讀不同 ->加入用戶偏好或是歷史紀錄
2. 遇到不常見或是新的 emoji ->定期更新訓練的 dataset
3. 中西方文化對 emoji 使用不同 ->擴大 dataset 的使用情境及面向

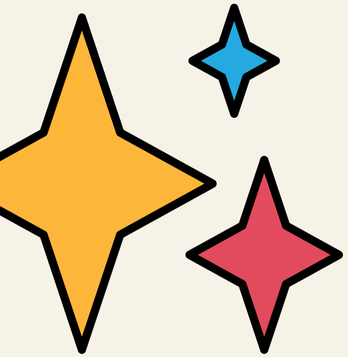
10. Conclusion

- BLSTM模型預測效果最佳。
- 使用更貼近中文實際使用情境的Instagram資料集。
- 最終模型的accuracy較baseline低，可能的原因是dataset多元性問題，如Twitter的資料是從最近七日推文爬取，而Instagram要限定帳號爬取資料。
- related works都只有20種emojis，而我們的模型能夠有效預測多達40種常用emoji。

Reference

- https://github.com/andylolu2/Text2Emoji
- https://www.kaggle.com/code/aguschin/lyrics-to-emoji
- https://github.com/Defcon27/Emoji-Prediction-using-DeepLearning/tree/master
- https://github.com/leyaoliatan/Sentiments-in-Tweets-with-Emojis/tree/main
- https://www.wcse.org/WCSE_2020_Summer/011.pdf
- https://huggingface.co/datasets/tweet_eval




Thank You!